Adaptive Gaussian inverse regression with partially unknown operator

Jan Johannes Maik Schwarz*

Université catholique de Louvain

February 28, 2013

This work deals with the ill-posed inverse problem of reconstructing a function f given implicitly as the solution of $g = Af$, where A is a compact linear operator with unknown singular values and known eigenfunctions. We observe the function g and the singular values of the operator subject to Gaussian white noise with respective noise levels ε and σ. We develop a minimax theory in terms of both noise levels and propose an orthogonal series estimator attaining the minimax rates. This estimator requires the optimal choice of a dimension parameter depending on certain characteristics of f and A. This work addresses the fully data-driven choice of the dimension parameter combining model selection with Lepski's method. We show that the fully data-driven estimator preserves minimax optimality over a wide range of classes for f and A and noise levels ε and σ. The results are illustrated considering Sobolev spaces and mildly and severely ill-posed inverse problems.

AMS 2010 subject classifications: 62G05, 62G08.

Keywords: Gaussian sequence space model, minimax theory, adaptive nonparametric estimation, model selection, Lepski's method, Sobolev spaces, mildly and severely ill-posed inverse problems

Acknowledgments This work was supported by the IAP research network no. P6/03 the Belgian Government (Belgian Science Policy) and by the "Fonds Spéciaux de Recherche" from the Université catholique de Louvain.

1. Introduction

Let $(H, \langle\cdot,\cdot\rangle_H)$ and $(G, \langle\cdot,\cdot\rangle_G)$ be separable Hilbert spaces and A a compact linear operator from H to G with unknown singular values. This work deals with the reconstruction of a function $f \in H$ given noisy observations of the image $g = Af$ on the one hand and of the unknown sequence of singular values $a = (a_j)_{j\in\mathbb{N}}$ on the other hand. In other words, we consider a statistical inverse problem with partially unknown operator. There is a vast literature on statistical inverse problems. For the case where the operator is fully known, the reader may refer to Johnstone and Silverman (1990), Mair and Ruymgaart (1996), Mathé and Pereverzev (2001), and Cavalier et al. (2002) and the references

*Corresponding author. Institut de statistique, biostatistique et sciences actuarielles, Voie du Roman Pays 20, boîte L1.04.01, B-1348 Louvain-la-Neuve, Belgique, e-mail: maik.schwarz@uclouvain.be

therein. A typical illustration of such a situation is a deconvolution problem (cf. Ermakov (1990), Stefanski and Carroll (1990), and Fan (1991) among many others). For a more detailed discussion and motivation of the case of a partially unknown operator which we consider in this work, we refer the reader to Cavalier and Hengartner (2005). Efromovich (1997) and Neumann (1997) consider such a setting in the particular context of a density deconvolution problem.

Let us describe in more detail the model we are going to consider. We suppose that A admits a singular value decomposition $(a_j, \varphi_j, \psi_j)_{j\in\mathbb{N}}$ as follows. Denote by A^* the adjoint operator of A. Then, A^*A is a compact operator on H with eigenvalues $(a_j^2)_{j\in\mathbb{N}}$ whose associated orthonormal basis of eigenfunctions $\{\varphi_j\}$ we suppose to be known. Analogously, the operator AA^* has eigenvalues $(a_j^2)_{j\in\mathbb{N}}$ and known orthonormal eigenfunctions $\psi_j = \|A\varphi_j\|_G^{-1} A\varphi_j$ in G. Projecting the inverse problem $g = Af$ on the eigenfunctions, we obtain the system of equations $[g]_j := \langle g, \psi_j\rangle_G = a_j\langle f, \varphi_j\rangle_H$ for $j \in \mathbb{N}$. As the operator A is compact, the sequence of singular values tends to zero and the inverse problem is called ill-posed.

The solution f is characterized by its coefficients $[f]_j := \langle f, \varphi_j\rangle_H$. Our objective is their estimation based on the following observations:

$$Y_j = [g]_j + \sqrt{\varepsilon}\,\xi_j = a_j[f]_j + \sqrt{\varepsilon}\xi_j \qquad \text{and} \qquad X_j = a_j + \sqrt{\sigma}\,\eta_j \qquad (j \in \mathbb{N}), \tag{1.1}$$

where the ξ_j, η_j are iid. standard normally distributed random variables and $\varepsilon, \sigma \in (0,1)$ are noise levels. Thus we represent the problem at hand as a hierarchical Gaussian sequence space model.

Of course f can only be reconstructed from such observations if all the a_j are non-zero which is the case if and only if the operator A is injective. We assume this from now on, which allows us to write $f = \sum_{j=1}^{\infty}[g]_j a_j^{-1}\varphi_j$. Hence, an orthogonal series estimator of f is a natural approach:

$$\widehat{f}_k := \sum_{j=1}^{k} \frac{Y_j}{X_j}\mathbf{1}_{[X_j^2 \geqslant \sigma]}\,\varphi_j.$$

The threshold using the indicator function accounts for the uncertainty caused by estimating the a_j by X_j. It corresponds to X_j's noise level as an estimator of a_j, which is a natural choice (cf. Neumann, 1997, p.310f.). Note that $\widehat{f}_k$ depends on a dimension parameter k whose choice essentially determines the estimation accuracy. Its optimal choice generally depends on both unknown sequences $([f]_j)$ and (a_j). Our purpose is to establish an adaptive estimation procedure for the function f which does not depend on these sequences. More precisely, assuming that the solution and the operator belong to given classes $f \in \mathcal{F}$ and $A \in \mathcal{A}$, respectively, we shall measure the accuracy of an estimator $\widetilde{f}$ of f by the maximal weighted risk $\mathcal{R}_\omega(\widetilde{f}, \mathcal{F}, \mathcal{A}) := \sup_{f\in\mathcal{F}}\sup_{A\in\mathcal{A}}\mathbb{E}\|\widetilde{f} - f\|_\omega^2$ defined with respect to some weighted norm $\|\cdot\|_\omega := \sum_{j\in\mathbb{N}}\omega_j|[\cdot]_j|^2$, where $\omega := (\omega_j)_{j\in\mathbb{N}}$ is a strictly positive weight sequences. This allows us to quantify the estimation accuracy in terms of the mean integrated square error (MISE) not only of f itself, but as well of its derivatives, for example. Given observations $Y = (Y_j)_{j\in\mathbb{N}}$ and $X = (X_j)_{j\in\mathbb{N}}$ with respective noise levels ε and σ according to (1.1), the minimax risk with respect to the classes $\mathcal{F}$ and $\mathcal{A}$ is then defined as $\mathcal{R}_\omega^*(\varepsilon, \sigma, \mathcal{F}, \mathcal{A}) := \inf_{\widetilde{f}}\mathcal{R}_\omega(\widetilde{f}, \mathcal{F}, \mathcal{A})$, where the infimum is taken over all possible estimators $\widetilde{f}$ of f. An estimator $\widehat{f}$ is said to attain the minimax rate or to be minimax optimal with respect to $\mathcal{F}$ and $\mathcal{A}$ if there is a constant $C > 0$ depending on the classes only such that $\mathcal{R}_\omega(\widehat{f}, \mathcal{F}, \mathcal{A}) \leqslant C\,\mathcal{R}_\omega^*(\varepsilon, \sigma, \mathcal{F}, \mathcal{A})$ for all $\varepsilon, \sigma \in (0,1)$. An estimation procedure which is fully data-driven and minimax optimal for a wide range of classes $\mathcal{F}$ and $\mathcal{A}$ is called *adaptive.*

In the next section, we show that for a wide range of classes $\mathcal{F}$ and $\mathcal{A}$ the orthogonal series estimator $\widehat{f}_{k_\varepsilon^*}$ attains the minimax rate for an optimal choice k_ε^* of the dimension parameter. We illustrate this result considering subsets of Sobolev spaces for $\mathcal{F}$ and distinguishing two types of operator classes $\mathcal{A}$ specifying the decay of the singular values: If (a_j) decays polynomially, the inverse problem is called mildly ill-posed and severely ill-posed if they decay exponentially. However, k_ε^* is chosen

subject to a classical variance-squared-bias trade-off and depends on properties of both classes $\mathcal{F}$ and $\mathcal{A}$ which are unknown in general.

The last section is devoted to the development of a data-driven choice $\widehat{k}$ of k, following the general model selection scheme (Barron et al., 1999, cf.). This methodology requires the careful choice of a contrast function and a penalty term. In this work, we will use a contrast function inspired by the work of Goldenshluger and Lepski (2011) who consider bandwidth selection for kernel estimators. Given a random sequence $(\widehat{\text{pen}}_k)_{k\geqslant 1}$ of penalties, a random set $\{1,\dots,\widehat{K}_{\varepsilon,\sigma}\}$ of admissible dimension parameters and the random sequence of contrasts

$$\widehat{\Psi}_k := \max_{k\leqslant j\leqslant \widehat{K}_{\varepsilon,\sigma}} \left\{ \|\widehat{f}_j - \widehat{f}_k\|_\omega^2 - \widehat{\text{pen}}_j \right\} \qquad (k\in\mathbb{N}). \tag{1.2}$$

The dimension parameter is selected as the minimizer[1] of a penalized contrast

$$\widehat{k} := \operatorname*{argmin}_{1\leqslant k\leqslant \widehat{K}} \left\{ \widehat{\Psi}_k + \widehat{\text{pen}}_k \right\}. \tag{1.3}$$

We assess the accuracy of the fully data-driven estimator $\widehat{f}_{\widehat{k}}$ deriving an upper bound for $\mathcal{R}_\omega(\widehat{f}_{\widehat{k}},\mathcal{F},\mathcal{A})$. Obviously this upper bound heavily depends the random sequence $(\widehat{\text{pen}}_k)$ and the random upper bound $\widehat{K}$. However, we construct these objects in such a way that the resulting fully data-driven estimator $\widehat{f}_{\widehat{k}}$ is minimax optimal over a wide range of classes and thus adaptive. The more technical proofs and some auxiliary results are deferred to the appendix.

Hoffmann and Reiss (2008) also study adaptive estimation in linear inverse problems, but their method is limited to mildly ill-posed inverse problems with known degree of ill-posedness. Also, the theoretical framework is quite different: they focus on sparse representations and therefore consider estimators based on wavelet thresholding and show their rate-optimality and adaptivity properties over Besov spaces with respect to the corresponding norms.

Adaptive estimation in a hierarchical Gaussian sequence space model has previously been considered by Cavalier and Hengartner (2005). Though, the authors restrict their investigation to the mildly ill-posed case and to noise levels satisfying $\sigma\leqslant\varepsilon$. The new approach presented in this paper has the advantage of not requiring such restrictions. On the contrary, the influence of the two noise levels on the estimation accuracy is characterized. Moreover, the estimator presented in this paper can attain optimal convergence rates independently of whether the underlying inverse problem is mildly or severely ill-posed, for example, even when $\varepsilon \ll \sigma$. This is an important feature in applications where the reduction of the noise level σ can be costly. In (satellite or medical) imaging, for example, the observation of the sequence X may correspond to calibration measurements. In order to achieve an adequately high precision of these measures as to reduce the noise level σ sufficiently, one might have to repeat expensive experiments. It is thus desirable to know how the estimator performs when σ exceeds ε.

2. Minimax

In this section we develop a minimax theory for Gaussian inverse regression with respect to the classes

$$\mathcal{F}_\gamma^r := \left\{ h\in H \;\middle|\; \sum_{j\in\mathbb{N}} \gamma_j |[h]_j|^2 =: \|h\|_\gamma^2 \leqslant r \right\} \text{ and}$$

$$\mathcal{A}_\lambda^d := \left\{ T\in C(H,G) \;\middle|\; \text{The eigenvalues } \{u_j\} \text{ of } T^*T \text{ satisfy } 1/d \leqslant \frac{u_j^2}{\lambda_j} \leqslant d \quad \forall\, j\in\mathbb{N} \right\},$$

[1] For a sequence $(b_k)_{k\in\mathbb{N}}$ attaining a minimal value on $N\subset\mathbb{N}$, let $\operatorname*{argmin}_{n\in N} b_n := \min\{n\in N \mid b_n\leqslant b_k \;\forall k\in N\}$.

where $C(H,G)$ denotes the set of all compact linear operators from H to G having $\{\varphi_j\}$ and $\{\psi_j\}$ as eigenfunctions, respectively. The minimal regularity conditions on the solution, the operator and the weighted norm $\|\cdot\|_\omega$ which we need in this section are summarized in the following assumption.

Assumption 2.1 *Let $\gamma := (\gamma_j)_{j\in\mathbb{N}}$, $\omega := (\omega_j)_{j\in\mathbb{N}}$ and $\lambda := (\lambda_j)_{j\in\mathbb{N}}$ be strictly positive sequences of weights with $\gamma_1 = \omega_1 = \lambda_1 = 1$ such that ω/γ and λ are non-increasing, respectively.*

Illustration 2.2 As an illustration of the results below, we will consider weight sequences $\gamma_j = j^{2p}$, for which $\mathcal{F}_\gamma^r$ is a Sobolev space of p-times differentiable functions if we consider the trigonometric basis in $H = L^2[0,1]$. As for the operator, we will distinguish the cases $\lambda_j = j^{-2b}$, referred to as *mildly ill-posed* (**[m]**) and $\lambda_j = \exp(-j^{2b})$, the *severely ill-posed* case (**[s]**). Concerning the weighted norm, we will consider sequences[2] $\omega_j \sim j^{2s}$, such that $\|f\|_\omega = \|f^{(s)}\|_{L^2}$ for all $f \in \mathcal{F}_\gamma^r$. We will assume that $b \geqslant 0$ and $p \geqslant s \geqslant 0$, such that Assumption 2.1 is satisfied.

The following result states lower risk bounds for the estimation of f and thus describes the complexity of the problem.

Theorem 2.3 *Suppose that we observe sequences Y and X according to the model (1.1). Consider sequences ω, γ, and λ satisfying Assumption 2.1. For all $\varepsilon, \sigma \in (0,1)$, define*

$$\rho_{k,\varepsilon} := \max\Big(\frac{\omega_k}{\gamma_k}, \sum_{j=1}^{k} \frac{\varepsilon\omega_j}{\lambda_j}\Big), \quad \chi_\varepsilon := \min_{k\in\mathbb{N}} \rho_{k,\varepsilon}, \quad k_\varepsilon^* := \operatorname*{argmin}_{k\in\mathbb{N}} \rho_{k,\varepsilon}, \quad \kappa_\sigma := \max_{k\in\mathbb{N}}\Big\{\frac{\omega_k}{\gamma_k}\min\Big(1, \frac{\sigma}{\lambda_k}\Big)\Big\}. \tag{2.1}$$

If $\eta := \inf_{n\in\mathbb{N}}\{\chi_\varepsilon^{-1}\min(\omega_{k_\varepsilon^*}\gamma_{k_\varepsilon^*}^{-1}, \sum_{l=1}^{k_\varepsilon^*}\varepsilon\omega_l(\lambda_l)^{-1})\} > 0$, then

$$\inf_{\widetilde{f}} \mathcal{R}_\omega(\widetilde{f}, \mathcal{F}_\gamma^r, \mathcal{A}_\lambda^d) \geqslant \frac{1}{4d}\min(\eta, r)\min(r, 1/(2d), (1-d^{-1/2})^2)\ \max(\chi_\varepsilon, \kappa_\sigma),$$

where the infimum is to be taken over all possible estimators $\widetilde{f}$ of f.

It is noteworthy that apart from the unwieldy constant, the lower bound is given by two terms (χ_ε and κ_σ), each of which depending only on one noise level. We show in the proof that χ_ε is actually, up to a constant, a lower risk bound uniformly for any known operator A in the class $\mathcal{A}_\lambda^d$. Hence, in this case no supremum over the class $\mathcal{A}_\lambda^d$ would be needed. The term κ_σ only arises if the operator is unknown in $\mathcal{A}_\lambda^d$. The proof of this lower bound is based on a comparison of different inverse problems with different operators in $\mathcal{A}_\lambda^d$, whence the supremum over this class. The term κ_σ quantifies to which extent the additional difficulty arising from the preliminary estimation of the eigenvalues a_j influences the possible estimation accuracy for f: As long as $\chi_\varepsilon \geqslant \kappa_\sigma$, the same lower bound as in the case of known eigenvalues holds. Otherwise, the lower bound increases. Notice further that the term $\rho_{k,\varepsilon}$ above corresponds to the MISE of the orthogonal series estimator $\widehat{f}_k$ in the case of known eigenvalues a_j, and k_ε^* is its minimizer with respect to k. Under classical smoothness assumptions, the rates and k_ε^* take the following forms.

Illustration 2.4 In the special cases defined in Illustration 2.2 above, the rates from (2.1) are

[m] $\quad \chi_\varepsilon \sim \varepsilon^{2(p-s)/(2p+2b+1)}, \quad k_\varepsilon^* \sim \varepsilon^{-1/(2p+2b+1)}, \quad \kappa_\sigma \sim \sigma^{((p-s)\wedge b)/b}$

[s] $\quad \chi_\varepsilon \sim |\log\varepsilon|^{(p-s)/b}, \quad k_\varepsilon^* \sim |\log\varepsilon|^{1/(2b)}, \quad \kappa_\sigma \sim |\log\sigma|^{-(p-s)/b}.$

The following theorem shows that the orthogonal series estimator $\widehat{f}_{k_\varepsilon^*}$ with optimal parameter k_ε^* given in (2.1) actually attains the lower risk bound up to a constant and is thus minimax optimal.

Theorem 2.5 *Under the assumptions of Theorem 2.3, the estimator $\widehat{f}_{k_\varepsilon^*}$ satisfies for all $\varepsilon, \sigma \in (0,1)$*

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{A\in\mathcal{A}_\lambda^d}\Big\{\mathbb{E}\|\widehat{f}_{k_\varepsilon^*} - f\|_\omega^2\Big\} \leqslant 4(6d+r)\ \max(\chi_\varepsilon, \kappa_\sigma).$$

[2] $b_\rho \sim c_\rho$ means that $\lim_{\rho\to 0} b_\rho/c_\rho$ exists in $(0,\infty)$.

To conclude this section, let us summarize the resulting optimal convergence rates under the classical smoothness assumptions introduced in Illustration 2.2. In order to characterize the influence of the second noise level σ, we consider it as a function of the first noise level ε.

Illustration 2.6 Let $(\sigma_\varepsilon)_{\varepsilon\in(0,1)}$ be a noise level in X depending on the noise level ε in Y.

[m] Let $p>1/2$, $b>1$, and $0\leqslant s\leqslant p$. If $q_1 := \lim_{\varepsilon\to 0}\varepsilon^{-2((p-s)\vee b)/(2p+2b+2)}\sigma_\varepsilon$ exists[3], then

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{A\in\mathcal{A}_\lambda^d}\mathbb{E}\|\widehat{f}^{(s)}_{k^*_\varepsilon}-f^{(s)}\|^2_{L^2}=\begin{cases}O(\varepsilon^{2(p-s)/(2p+2b+1)}) & \text{if } q_1<\infty\\ O(\sigma_\varepsilon^{((p-s)\wedge b)/b}) & \text{otherwise.}\end{cases}$$

[s] Let $p>1/2$,$b>0$ and $0\leqslant s\leqslant p$. If $q_2 := \lim_{\varepsilon\to 0}|\log\varepsilon|\,|\log\sigma_\varepsilon|^{-1}$ exists, then

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{A\in\mathcal{A}_\lambda^d}\mathbb{E}\|\widehat{f}^{(s)}_{k^*_\varepsilon}-f^{(s)}\|^2_{L^2}=\begin{cases}O(|\log\varepsilon|^{(p-s)/b}) & \text{if } q_2<\infty\\ O(|\log\sigma_\varepsilon|^{(p-s)/b}) & \text{otherwise.}\end{cases}$$

This illustration shows that often the same optimal rates as in the case of known eigenvalues hold even when $\varepsilon<\sigma$.

3. Adaptation

In this section, we construct a fully data-driven estimator of f following the procedure sketched in (1.2) and (1.3). The following Lemma will be our key tool when controlling the risk of the adaptive estimator.

Lemma 3.1 *Let* pen *be an arbitrary positive sequence and* $K\in\mathbb{N}$. *Consider the sequence* Ψ *of contrasts* $\Psi_k := \max_{k\leqslant j\leqslant K}\left\{\|\widehat{f}_j-\widehat{f}_k\|^2_\omega-\mathrm{pen}_j\right\}$ *and* $\widetilde{k} := \operatorname{argmin}_{1\leqslant j\leqslant K}\{\Psi_j+\mathrm{pen}_j\}$. *Let further* $(t)_+ := (t\vee 0)$. *If* $(\mathrm{pen}_1,\dots,\mathrm{pen}_K)$ *is non-decreasing, then we have for all* $1\leqslant k\leqslant K$ *that*

$$\|\widehat{f}_{\widetilde{k}}-f\|^2_\omega\leqslant 7\,\mathrm{pen}_k+78\,\mathrm{bias}^2_k+42\max_{1\leqslant j\leqslant K}\left(\|\widehat{f}_j-f_j\|^2_\omega-\frac{1}{6}\,\mathrm{pen}_j\right)_+,\tag{3.1}$$

where we denote by $f_j := \sum_{k=1}^j[f]_k\,\varphi_k$ *the projection of* f *on the first* j *basis vectors in* H *and by* $\mathrm{bias}_k := \sup_{j\geqslant k}\|f-f_j\|_\omega$ *the bias due to the projection.*

Proof. In view of the definition of $\widetilde{k}$, we have for all $1\leqslant k\leqslant K$ that

$$\begin{aligned}\|\widehat{f}_{\widetilde{k}}-f\|^2_\omega&\leqslant 3\Big\{\|\widehat{f}_{\widetilde{k}}-\widehat{f}_{k\wedge\widetilde{k}}\|^2_\omega+\|\widehat{f}_{k\wedge\widetilde{k}}-\widehat{f}_k\|^2_\omega+\|\widehat{f}_k-f\|^2_\omega\Big\}\\&\leqslant 3\Big\{\Psi_k+\mathrm{pen}_{\widetilde{k}}+\Psi_{\widetilde{k}}+\mathrm{pen}_k+\|\widehat{f}_k-f\|^2_\omega\Big\}\\&\leqslant 6\Big\{\Psi_k+\mathrm{pen}_k\Big\}+3\|\widehat{f}_k-f\|^2_\omega\end{aligned}\tag{3.2}$$

Since $(\mathrm{pen}_1,\dots,\mathrm{pen}_K)$ is non-decreasing and $4\,\mathrm{bias}^2_k\geqslant\max_{k\leqslant j\leqslant K}\|f_k-f_j\|^2_\omega$, we have

$$\Psi_k\leqslant 6\max_{1\leqslant j\leqslant K}\left(\|\widehat{f}_j-f_j\|^2_\omega-\frac{1}{6}\,\mathrm{pen}_j\right)_++12\,\mathrm{bias}^2_k\,.$$

It easily verified that for all $1\leqslant k\leqslant K$ we have

$$\|\widehat{f}_k-f\|^2_\omega\leqslant\frac{1}{3}\,\mathrm{pen}_k+2\,\mathrm{bias}^2_k+2\max_{1\leqslant j\leqslant K}\left(\|\widehat{f}_j-f_j\|^2_\omega-\frac{1}{6}\,\mathrm{pen}_j\right)_+.$$

[3]The limit «∞» meaning strict divergence is authorized.

The result follows combining the last estimates with (3.2). $\square$

The Lemma being valid for any upper bound K and any monotonic sequence of penalties pen, we need to specify our choice. Let us first define some auxiliary quantities required in the construction of the random penalty sequence $\widehat{\text{pen}}$ and the upper bound $\widehat{K}$.

Definition 3.2 *For any sequence $\alpha := (\alpha_j)_{j\in\mathbb{N}}$, define*

(i) $\Delta_k^\alpha := \max_{1\leqslant j\leqslant k} \omega_j\, \alpha_j^{-2}$ *and* $\delta_k^\alpha := k\Delta_k^\alpha \frac{\log(\Delta_k^\alpha \vee (k+2))}{\log(k+2)}$;

(ii) given $\omega_k^+ := \max_{1\leqslant j\leqslant k} \omega_j$, $N_\varepsilon^\circ := \max\{1 \leqslant N \leqslant \varepsilon^{-1} \mid \omega_N^+ \leqslant \varepsilon^{-1}\}$, *and* $v_\sigma := (8\log(\log(\sigma^{-1}+20)))^{-1}$, *let*

$$N_\varepsilon^\alpha := \min\Big\{2 \leqslant j \leqslant N_\varepsilon^\circ \;\Big|\; \frac{\alpha_j^2}{j\omega_j^+} \leqslant \varepsilon|\log\varepsilon|\Big\} - 1 \quad \text{and} \quad M_\sigma^\alpha := \min\Big\{2 \leqslant j \leqslant \sigma^{-1} \;\Big|\; \alpha_j^2 \leqslant \sigma^{1-v_\sigma}\Big\} - 1,$$

and $K_{\varepsilon,\sigma}^\alpha := N_\varepsilon^\alpha \wedge M_\sigma^\alpha$. *If the defining set is empty, set* $N_\varepsilon^\alpha = N_\varepsilon^\circ$ *or* $M_\sigma^\alpha = \lfloor \sigma^{-1} \rfloor$, *respectively.*

Choosing appropriate sequences α, these quantities allow us define the random penalty term needed for the data-driven choice of k as well as its deterministic counterpart which will be used in the control of the risk.
Using this definition and denoting by X the sequence of random variables $(X_j)_{j\in\mathbb{N}}$, define

$$\widehat{K}_{\varepsilon,\sigma} := K_{\varepsilon,\sigma}^X \quad \text{and} \quad \widehat{\text{pen}}_k := 600\delta_k^X\,\varepsilon. \tag{3.3}$$

Substituting these definitions in (1.2) and (1.3) yields a choice of the dimension parameter k depending exclusively on the observations and the noise levels, but not on any underlying smoothness classes.
Consider the upper risk bound in Lemma 3.1. In order to control the risk of the data-driven estimator, we decompose it with respect to an event on which the randomized quantities $\widehat{\text{pen}}_k$ and $\widehat{K}_{\varepsilon,\sigma}$ are close to some deterministic counterparts pen_k^a, $K_{\varepsilon,\sigma}^-$, and $K_{\varepsilon,\sigma}^+$ to be defined below in Propositions 3.3 and 3.5. More precisely, consider the event

$$\mho_{\varepsilon,\sigma} := \{\text{pen}_k^a \leqslant \widehat{\text{pen}}_k \leqslant 30\,\text{pen}_k^a \quad \forall\, 1 \leqslant k \leqslant K_{\varepsilon,\sigma}^+\} \cap \{K_{\varepsilon,\sigma}^- \leqslant \widehat{K}_{\varepsilon,\sigma} \leqslant K_{\varepsilon,\sigma}^+\}$$

and the corresponding risk decomposition

$$\mathbb{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 = \mathbb{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}_{\mho_{\varepsilon,\sigma}} + \mathbb{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}_{\mho_{\varepsilon,\sigma}^c}. \tag{3.4}$$

As the random sequence $\widehat{\text{pen}}_k$ is non-decreasing in k by construction, we may apply Lemma 3.1 and obtain for every $1 \leqslant k \leqslant \widehat{K}_{\varepsilon,\sigma}$

$$\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \leqslant 7\,\widehat{\text{pen}}_k + 78\,\text{bias}_k^2 + 42 \max_{1\leqslant j\leqslant \widehat{K}_{\varepsilon,\sigma}} \Big(\|\widehat{f}_j - f_j\|_\omega^2 - \frac{1}{6}\widehat{\text{pen}}_j\Big)_+.$$

On the event $\mho_{\varepsilon,\sigma}$, this implies that

$$\mathbb{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}_{\mho_{\varepsilon,\sigma}} \leqslant 420 \min_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^-} \{\max(\text{pen}_k^a, \text{bias}_k^2)\} + 42 \max_{1\leqslant j\leqslant K_{\varepsilon,\sigma}^+} \mathbb{E}\Big(\|\widehat{f}_j - f_j\|_\omega^2 - \frac{1}{6}\,\text{pen}_j^a\Big)_+. \tag{3.5}$$

The second term in the last inequality is controlled uniformly over $\mathcal{F}_\gamma^r$ and $\mathcal{A}_\lambda^d$ by the following Proposition.

Proposition 3.3 *Given $A \in \mathcal{A}_\lambda^d$ with singular values $a := (a_j)_{j\in\mathbb{N}}$, let $\sqrt{4d\lambda} := (\sqrt{4d\lambda_j})_{j\in\mathbb{N}}$ and define $K_{\varepsilon,\sigma}^+ := K_{\varepsilon,\sigma}^{\sqrt{4d\lambda}}$, $M_{\varepsilon,\sigma}^+ := M_{\varepsilon,\sigma}^{\sqrt{4d\lambda}}$, and $\mathrm{pen}_k^a := 60\delta_k^a\,\varepsilon$ using Definition 3.2. There is a constant $C > 0$ depending only on the class $\mathcal{A}_\lambda^d$ such that*

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{A\in\mathcal{A}_\lambda^d}\mathbb{E}\Big[\max_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^+}\Big(\|\widehat{f}_k - f_k\|_\omega^2 - \frac{1}{6}\,\mathrm{pen}_k^a\Big)_+\Big] \leqslant C\,\Big\{\varepsilon + r\kappa_\sigma + \sigma\Big\}.$$

Roughly speaking, the penalty term is an upper bound for the estimator's variation. Typically, it can be chosen as a multiple of the estimator's variance. Thus, inequality (3.1) actually features a bias variance decomposition of the risk with an additional third term which is controlled by the above proposition.

Illustration 3.4 Note that for any operator $A \in \mathcal{A}_\lambda^d$ with sequence $(a_j)_{j\geqslant 1}$ of singular values, the sequence δ^a appearing in the definition of the penalty term pen^a satisfies $(d\zeta_d)^{-1} \leqslant (\delta_j^a/\delta_j^\lambda) \leqslant d\zeta_d$ for all $j \in \mathbb{N}$, with $\zeta_d = \log(3d)/\log(3)$. In the special cases defined in Illustration 2.2 above, the sequence δ^λ takes the following form:

[m] $\delta_k^\lambda \sim k^{2b+2s+1}$ **[s]** $\delta_k^\lambda \sim k^{2b+2s+1}\exp(k^{2b})(\log k)^{-1}$.

The next proposition ensures that the randomized upper bound and penalty sequence behave similarly to their deterministic counterparts with sufficiently high probability so as not to deteriorate the estimation risk. In view of Proposition 3.3, this justifies the choice of the penalty.

Proposition 3.5 *Let $K_{\varepsilon,\sigma}^- := K_{\varepsilon,\sigma}^{\sqrt{\lambda/(4d)}}$ and $M_\sigma^+ := M_\sigma^{\sqrt{4d\lambda}}$ using Definition 3.2 and suppose that there is a constant $L > 0$ depending only on λ and d such that*

$$\sigma^{-7}\lambda_{M_\sigma^+ +1}^{-1/2}\exp\Big(-\lambda_{M_\sigma^+ +1}/(72\,\sigma d)\Big) \leqslant L \quad \text{for all} \quad \sigma \in (0,1). \tag{3.6}$$

Then, there is a constant $C > 0$ depending only on the class $\mathcal{A}_\lambda^d$ such that

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{A\in\mathcal{A}_\lambda^d}\mathbb{E}[\|\widehat{f}_{\widehat{k}} - f\|_\omega^2\mathbf{1}_{\mho_{\varepsilon,\sigma}^c}] \leqslant C\,(1+r)\,\sigma \quad \text{for all} \quad \varepsilon,\sigma \in (0,1).$$

Condition (3.6) is satisfied in particular under the classical smoothness assumptions considered in the illustrations. We are finally prepared to state the upper risk bound of the fully data-driven estimator $\widehat{f}_{\widehat{k}}$ of f, which is the main result of this article.

Theorem 3.6 *Under Assumption 2.1 and supposing (3.6), there is a constant C depending only on the class $\mathcal{A}_\lambda^d$ such that for all $\varepsilon,\sigma \in (0,1)$ the adaptive estimator $\widehat{f}_{\widehat{k}}$ satisfies*

$$\mathcal{R}_\omega(\widehat{f}_{\widehat{k}}, \mathcal{F}_\gamma^r, \mathcal{A}_\lambda^d) \leqslant C\,(1+r)\Big\{\min_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^-}\{\max(\omega_k/\gamma_k, \delta_k^\lambda\varepsilon)\} + \kappa_\sigma + \varepsilon + \sigma\Big\}.$$

Proof. Considering (3.5), note that for all $A \in \mathcal{A}_\lambda^d$, we have $\mathrm{pen}_k^a \leqslant 60\varepsilon d\zeta_d\delta_k^\lambda$ with $\zeta_d = \log(3d)/\log(3)$. On the other hand, it is easily seen that for all $f \in \mathcal{F}_\gamma^r$, one has $\mathrm{bias}_k^2 \leqslant r\,(\omega_k/\gamma_k)$. Thus, we can write

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{A\in\mathcal{A}_\lambda^d}\min_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^-}\{\max(\mathrm{pen}_k^a, \mathrm{bias}_k^2)\} \leqslant C\,(1+r)\min_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^-}\{\max(\omega_k/\gamma_k, \delta_k^\lambda\varepsilon)\}$$

for some constant $C > 0$ depending only on d. In view of (3.4), the rest of the proof is obvious using Propositions 3.3 and 3.5. $\square$

A comparison with the lower bound from Theorem 2.3 shows that this upper bound ensures minimax optimality of the adaptive estimator $\widehat{f}_{\widehat{k}}$ only if

$$\chi_{\varepsilon,\sigma}^\diamond := \min_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^-}\Big[\max\Big(\frac{\omega_k}{\gamma_k}, \delta_k^\lambda\varepsilon\Big)\Big]$$

is at most of the same order as $\max(\chi_\varepsilon, \kappa_\sigma)$, whence the following corollary.

Corollary 3.7 *Under Assumption 2.1 and if* $\sup_{\varepsilon,\sigma\in(0,1)}\{\chi^{\diamond}_{\varepsilon,\sigma}/\max(\chi_\varepsilon,\kappa_\sigma)\}<\infty$*, we have*

$$\mathcal{R}_\omega(\widehat{f}_{\widehat{k}},\mathcal{F}^r_\gamma,\mathcal{A}^d_\lambda)\leqslant C\,\mathcal{R}^*_\omega(\mathcal{F}^r_\gamma,\mathcal{A}^d_\lambda)\qquad\forall\,\varepsilon,\sigma\in(0,1).$$

We conclude this article reconsidering the framework of the preceding Illustration 2.6. Notice that the adaptive estimator is minimax optimal over a wide range of cases, even when $\varepsilon<\sigma$.

Illustration 3.8 Let $(\sigma_\varepsilon)_{\varepsilon\in(0,1)}$ be a noise level in X depending on the noise level ε in Y and suppose that the limits q_1 and q_2 from Illustration 2.6 exist in the respective cases. Some straightforward computations then show that the adaptive estimator attains the following rates of convergence.

[m] If $p-s>b$, the adaptive estimator $f^{(s)}_{\widehat{k}}$ attains the optimal rates (cf. Illustration 2.6). In case $p-s\leqslant b$, we have, supposing that $q^v_1:=\lim\limits_{\varepsilon\to0}\varepsilon^{-2b/(2p+2b+1)}\sigma_\varepsilon^{1-v_{\sigma_\varepsilon}}$ exists,

$$\sup_{f\in\mathcal{F}^r_\gamma}\sup_{A\in\mathcal{A}^d_\lambda}\mathbb{E}\|\widehat{f}^{(s)}_{\widehat{k}}-f^{(s)}\|^2_{L^2}=\begin{cases}O(\varepsilon^{2(p-s)/(2p+2b+1)}) & \text{if } q_1<\infty \text{ and } q^v_1<\infty,\\ O(\sigma_\varepsilon^{(p-s)/b}\sigma_\varepsilon^{-v_{\sigma_\varepsilon}}) & \text{otherwise.}\end{cases}$$

[s] The adaptive estimator attains the optimal rates.

A. Proofs

A.1. Minimax theory (Section 2)

Lower risk bound

Proof of Theorem 2.3. The proof consists of two steps: (A) First, we show that χ_ε yields a lower risk bound in the case where the eigenvalues (a_j) of the operator A are known. (B) Then, we show that another lower risk bound is given by κ_σ.
Step (A). Given $\zeta:=\eta\min(r,1/(2d))$ and $\alpha_\varepsilon:=\chi_\varepsilon(\sum_{j=1}^{k^*_\varepsilon}\varepsilon\omega_j/\lambda_j)^{-1}$ we consider the function $f:=(\varepsilon\zeta\alpha_\varepsilon)^{1/2}\sum_{j=1}^{k^*_\varepsilon}\lambda_j^{-1/2}\varphi_j$. We are going to show that for any $\theta:=(\theta_j)\in\{-1,1\}^{k^*_\varepsilon}$, the function $f_\theta:=\sum_{j=1}^{k^*_\varepsilon}\theta_j[f]_j\varphi_j$ belongs to $\mathcal{F}^r_\gamma$ and is hence a possible candidate for the solution.
For a fixed θ and under the hypothesis that the solution is f_θ, the observation Y_k is distributed according to $\mathcal{N}(a_k[f_\theta]_k,\varepsilon)$ for any $k\in\mathbb{N}$. We denote by $\mathbb{P}_\theta$ the distribution of the resulting sequence $\{Y_k\}$ and by $\mathbb{E}_\theta$ the expectation with respect to this distribution.
Furthermore, for $1\leqslant j\leqslant k^*_\varepsilon$ and each θ, we introduce $\theta^{(j)}$ by $\theta^{(j)}_l=\theta_l$ for $j\neq l$ and $\theta^{(j)}_j=-\theta_j$. The key argument of this proof is the following reduction scheme. If $\widetilde{f}$ denotes an estimator of f then we conclude

$$\begin{aligned}\sup_{f\in\mathcal{F}^r_\gamma}\mathbb{E}\|\widetilde{f}-f\|^2_\omega&\geqslant\sup_{\theta\in\{-1,1\}^{k^*_\varepsilon}}\mathbb{E}_\theta\|\widetilde{f}-f_\theta\|^2_\omega\geqslant\frac{1}{2^{k^*_\varepsilon}}\sum_{\theta\in\{-1,1\}^{2k^*_\varepsilon}}\mathbb{E}_\theta\|\widetilde{f}-f_\theta\|^2_\omega\\&\geqslant\frac{1}{2^{k^*_\varepsilon}}\sum_{\theta\in\{-1,1\}^{k^*_\varepsilon}}\sum_{j=1}^{k^*_\varepsilon}\omega_j\mathbb{E}_\theta|[\widetilde{f}-f_\theta]_j|^2\\&=\frac{1}{2^{k^*_\varepsilon}}\sum_{\theta\in\{-1,1\}^{k^*_\varepsilon}}\sum_{j=1}^{k^*_\varepsilon}\frac{\omega_j}{2}\Big\{\mathbb{E}_\theta|[\widetilde{f}-f_\theta]_j|^2+\mathbb{E}_{\theta^{(j)}}|[\widetilde{f}-f_{\theta^{(j)}}]_j|^2\Big\}.\end{aligned}\tag{A.1}$$

Below we show furthermore that for all $\varepsilon\in(0,1)$ we have

$$\Big\{\mathbb{E}_\theta|[\widetilde{f}-f_\theta]_j|^2+\mathbb{E}_{\theta^{(j)}}|[\widetilde{f}-f_{\theta^{(j)}}]_j|^2\Big\}\geqslant\frac{\varepsilon\,\zeta\alpha_\varepsilon}{2\lambda_j}.\tag{A.2}$$

Combining the last lower bound and the reduction scheme gives

$$\sup_{f\in\mathcal{F}_\gamma^r}\mathbb{E}\|\widetilde{f}-f\|_\omega^2 \geqslant \frac{1}{2^{k_\varepsilon^*}}\sum_{\theta\in\{-1,1\}^{k_\varepsilon^*}}\sum_{j=1}^{k_\varepsilon^*}\frac{\omega_j}{2}\frac{\varepsilon\zeta\alpha_\varepsilon}{2\lambda_j} = \frac{\zeta\alpha_\varepsilon}{4}\sum_{j=1}^{k_\varepsilon^*}\frac{\varepsilon\omega_j}{\lambda_j} = \frac{\zeta\chi_\varepsilon}{4},$$

which implies the lower bound given in the theorem by definition of ζ.
To complete the proof, it remains to check (A.2) and $f_\theta \in \mathcal{F}_\gamma^r$ for all $\theta\in\{-1,1\}^{k_\varepsilon^*}$. The latter is easily verified if $f\in\mathcal{F}_\gamma^r$, which can be seen recalling that ω/γ is non-increasing and noticing that the definitions of ζ, α_ε and η imply $\|f\|_\gamma^2 \leqslant \zeta\frac{\gamma_{k_\varepsilon^*}}{\omega_{k_\varepsilon^*}}\alpha_\varepsilon\Big(\sum_{j=1}^{k_\varepsilon^*}\frac{\varepsilon\omega_j}{\lambda_j}\Big)\leqslant \zeta/\eta\leqslant r$.
It remains to show (A.2). Consider the Hellinger affinity $\rho(\mathbb{P}_1,\mathbb{P}_{-1})=\int\sqrt{d\mathbb{P}_1\, d\mathbb{P}_{-1}}$, then we obtain for any estimator $\widetilde{f}$ of f that

$$\begin{aligned}\rho(\mathbb{P}_1,\mathbb{P}_{-1}) &\leqslant \int\frac{|[\widetilde{f}-f_{\theta^{(j)}}]_j|}{|[f_\theta-f_{\theta^{(j)}}]_j|}\sqrt{d\mathbb{P}_1\, d\mathbb{P}_{-1}}+\int\frac{|[\widetilde{f}-f_{\theta}]_j|}{|[f_\theta-f_{\theta^{(j)}}]_j|}\sqrt{d\mathbb{P}_1\, d\mathbb{P}_{-1}}\\ &\leqslant \Big(\int\frac{|[\widetilde{f}-f_{\theta^{(j)}}]_j|^2}{|[f_\theta-f_{\theta^{(j)}}]_j|^2}d\mathbb{P}_1\Big)^{1/2}+\Big(\int\frac{|[\widetilde{f}-f_{\theta}]_j|^2}{|[f_\theta-f_{\theta^{(j)}}]_j|^2}d\mathbb{P}_{-1}\Big)^{1/2}.\end{aligned}$$

Rewriting the last estimate we obtain

$$\Big\{\mathbb{E}_\theta|[\widetilde{f}-f_\theta]_j|^2+\mathbb{E}_{\theta^{(j)}}|[\widetilde{f}-f_{\theta^{(j)}}]_j|^2\Big\}\geqslant\frac{1}{2}|[f_\theta-f_{\theta^{(j)}}]_j|^2\rho^2(\mathbb{P}_1,\mathbb{P}_{-1}). \tag{A.3}$$

Next, we bound the Hellinger affinity $\rho(\mathbb{P}_1,\mathbb{P}_{-1})$ from below. Consider the Kullback-Leibler divergence of these two distributions first. The components of the two sequences corresponding to the distributions $\mathbb{P}_1$ and $\mathbb{P}_{-1}$ are pairwise equally distributed except for the j-th component. Thus, we have $\log(d\mathbb{P}_\theta/d\mathbb{P}_{\theta^{(j)}})=(2y_ja_j\theta_j[f]_j/\varepsilon)$, and taking the integral over y_j with respect to $\mathbb{P}_\theta$, we find

$$KL(\mathbb{P}_1,\mathbb{P}_{-1})=\frac{2}{\varepsilon}\,a_j^2[f]_j^2\leqslant\frac{2d}{\varepsilon}\,[f]_j^2\lambda_j=2d\zeta\alpha_\varepsilon\leqslant 1,$$

Using the well-known relationship $\rho(\mathbb{P}_1,\mathbb{P}_{-1})\geqslant 1-(1/2)KL(\mathbb{P}_1,\mathbb{P}_{-1})$ between the Kullback-Leibler divergence and the Hellinger affinity, we obtain that $\rho(\mathbb{P}_1,\mathbb{P}_{-1})\geqslant 1/2$. Using this estimate, (A.3) becomes $\Big\{\mathbb{E}_\theta|[\widetilde{f}-f_\theta]_j|^2+\mathbb{E}_{\theta^{(j)}}|[\widetilde{f}-f_{\theta^{(j)}}]_j|^2\Big\}\geqslant\frac{1}{2}[f]_j^2$, and combining this with (A.1) implies the result by construction of the solution f.

Step (B). First, we construct two solutions $f_\theta\in\mathcal{F}_\gamma^r$ and operators $A_\theta\in\mathcal{A}_\lambda^d$ (with $\theta\in\{-1,1\}$) such that the resulting images g_θ satisfy $g_{-1}=g_1$. To this end, we define $k_\sigma^*:=\operatorname{argmax}_{j\in\mathbb{N}}\{\omega_j\gamma_j^{-1}\min(1,\sigma\lambda_j^{-1})\}$ and $\alpha_\sigma:=\zeta\min(1,\sigma^{1/2}\lambda_{k_\sigma^*}^{-1/2})$ with $\zeta:=\min(2^{-1},(1-d^{-1/2}))$. Observe that $1\geqslant(1-\alpha_\sigma)^2\geqslant(1-(1-1/d^{1/2}))^2\geqslant 1/d$ and $1\leqslant(1+\alpha_\sigma)^2\leqslant(1+(1-1/d^{1/2}))^2=(2-1/d^{1/2})^2\leqslant d$, which implies $1/d\leqslant(1+\theta\alpha_\sigma)^2\leqslant d$. These inequalities will be used below without further reference. We show below that for each θ the function $f_\theta:=(1-\theta\alpha_\sigma)\frac{r}{d}\gamma_{k_\sigma^*}^{-1/2}\varphi_{k_\sigma^*}$ belongs to $\mathcal{F}_\gamma^r$ and that the operator A_θ with the singular values $a_k^\theta=[1+\theta\alpha_\sigma\mathbf{1}\{k=k_\sigma^*\}]\,\sqrt{\lambda_k}$ is an element of $\mathcal{A}_\lambda^d$. We obviously have that $A_1f_f=(1-\alpha_\sigma^2)(\lambda_{k_\sigma^*}/\gamma_{k_\sigma^*})^{1/2}(r/d)\psi_{k_\sigma^*}=A_{-1}f_{-1}$.
For $\theta\in\{\pm1\}$, denote by $\mathbb{P}_\theta$ the joint distribution of the two sequences $(X_1,X_2,\dots)$ and $(Y_1,Y_2,\dots)$, and let $\mathbb{E}_\theta$ denote the expectation with respect to $\mathbb{P}_\theta$.
Applying a reduction scheme as under Step (A) above, we deduce that for each estimator $\widetilde{f}$ of f

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{A\in\mathcal{A}_\lambda^d}\mathbb{E}\|\widetilde{f}-f\|_\omega^2\geqslant\max_{\theta\in\{-1,1\}}\mathbb{E}_\theta\|\widetilde{f}-f_\theta\|_\omega^2\geqslant\frac{1}{2}\Big\{\mathbb{E}_1\|\widetilde{f}-f_1\|_\omega^2+\mathbb{E}_{-1}\|\widetilde{f}-f_{-1}\|_\omega^2\Big\}.$$

Below we show furthermore that

$$\mathbb{E}_1\|\widetilde{f}-f_1\|_\omega^2+\mathbb{E}_{-1}\|\widetilde{f}-f_{-1}\|_\omega^2 \geqslant \frac{1}{8}\|f_1-f_{-1}\|_\omega^2. \tag{A.4}$$

Moreover, we have $\|f_1-f_{-1}\|_\omega^2 = 4\alpha_\sigma^2(r/d)\omega_{k_\sigma^*}\gamma_{k_\sigma^*}^{-1} = 4\zeta^2(r/d)\omega_{k_\sigma^*}\gamma_{k_\sigma^*}^{-1}\min\Big(1,\frac{\sigma}{\lambda_{k_\sigma^*}}\Big)$. Combining the last lower bound with the reduction scheme and the definition of k_σ^* implies the result of the theorem. To conclude the proof, it remains to check (A.4), $f_\theta\in\mathcal{F}_\gamma^r$ and $A_\theta\in\mathcal{A}_\lambda^d$ for both θ. In order to show $f_\theta\in\mathcal{F}_\gamma^r$, observe that $\|f_\theta\|_\gamma^2=\gamma_{k_\sigma^*}|[f_\theta]_{k_\sigma^*}|^2\leqslant\gamma_{k_\sigma^*}|(1-\theta\alpha_\sigma)(r/d)\gamma_{k_\sigma^*}^{-1/2}|^2\leqslant r$.
To check that $A_\theta\in\mathcal{A}_\lambda^d$, it remains to show that $1/d\leqslant(a_j^\theta)^2/\lambda_j\leqslant d$ for all $j\geqslant 1$. These inequalities are obviously satisfied for all $j\neq k_\sigma^*$, and as well for $j=k_\sigma^*$ by construction of the operator A. Finally consider (A.4). As in Step (A) above by employing the Hellinger affinity $\rho(\mathbb{P}_1,\mathbb{P}_{-1})$ we obtain for any estimator $\widetilde{f}$ of f that

$$\mathbb{E}_1\|\widetilde{f}-f_1\|_\omega^2+\mathbb{E}_{-1}\|\widetilde{f}-f_{-1}\|_\omega^2 \geqslant \frac{1}{2}\|f_1-f_{-1}\|_\omega^2\rho^2(\mathbb{P}_1,\mathbb{P}_{-1}).$$

Next, we bound the Hellinger affinity $\rho(\mathbb{P}_1,\mathbb{P}_{-1})$ from below for all $\sigma\in(0,1)$, which proves (A.4). Notice that by construction of f_θ and A_θ, the distribution of X_i and Y_i does not depend on θ, except for $X_{k_\sigma^*}^\theta$. It is thus easily seen that the Kullback-Leibler divergence can be controlled as follows,

$$KL(\mathbb{P}_1,\mathbb{P}_{-1})=\frac{(a_{k_\sigma^*}^1-a_{k_\sigma^*}^{-1})^2}{2\sigma}=\frac{2\alpha_\sigma^2}{\sigma}\lambda_{k_\sigma^*}\leqslant 1$$

Using $\rho(\mathbb{P}_1,\mathbb{P}_{-1})\geqslant 1-(1/2)KL(\mathbb{P}_1,\mathbb{P}_{-1})$ again, (A.4) is shown and so is the theorem. □

Upper risk bound

The following proof uses Lemma A.1 from the auxiliary results section A.3 below.
Proof of Theorem 2.5. Define $\widetilde{f}:=\sum_{j=1}^{k_\varepsilon^*}[f]_j\mathbf{1}\{X_j^2\geqslant\sigma\}e_j$ and decompose the risk into two terms,

$$\mathbb{E}\|\widehat{f}-f\|_\omega^2=\mathbb{E}\|\widehat{f}-\widetilde{f}\|_\omega^2+\mathbb{E}\|\widetilde{f}-f\|_\omega^2=:A+B, \tag{A.5}$$

which we bound separately. Consider first A which we decompose further,

$$\mathbb{E}\|\widehat{f}-\widetilde{f}\|_\omega^2=\sum_{j=1}^{k_\varepsilon^*}\omega_j\mathbb{E}\left[\frac{(Y_j-\mathbb{E}Y_j)^2}{X_j^2}\mathbf{1}\{X_j^2\geqslant\sigma\}\right] + \sum_{j=1}^{k_\varepsilon^*}\omega_j|[f]_j|^2\mathbb{E}\left[\frac{(X_j-\mathbb{E}X_j)^2}{X_j^2}\mathbf{1}\{X_j^2\geqslant\sigma\}\right]=:A_1+A_2.$$

As far as A_1 is considered, we use Lemma A.1 (iii) from Section A.3 below and write

$$A_1=\sum_{j=1}^{k_\varepsilon^*}\frac{\omega_j\varepsilon}{\mathbb{E}[X_j]^2}\,\mathbb{E}\left[\left(\frac{\mathbb{E}[X_j]}{X_j}\right)^2\mathbf{1}\{X_j^2\geqslant\sigma\}\right]\leqslant 4d\sum_{j=1}^{k_\varepsilon^*}\frac{\omega_j\varepsilon}{\lambda_j}\leqslant 4d\chi_\varepsilon.$$

As for A_2, we apply Lemma A.1 (i) and obtain

$$A_2\leqslant 8d\sum_{j=1}^{k_\varepsilon^*}\omega_j|[f]_j|^2\min\left(1,\frac{\sigma}{\lambda_j}\right)\leqslant 8d\kappa_\sigma$$

Consider now B which we decompose further into

$$\mathbb{E}\|\widetilde{f}-f\|_\omega^2 = \sum_{j\in\mathbb{N}} \omega_j|[f]_j|^2 \mathbb{E}[(1-\mathbf{1}\{1\leqslant j\leqslant k_\varepsilon^*\}\mathbf{1}\{X_j^2\geqslant\sigma\})^2]$$

$$= \sum_{j>k_\varepsilon^*}\omega_j|[f]_j|^2 + \sum_{j=1}^{k_\varepsilon^*}\omega_j|[f]_j|^2\mathbf{P}\Big(X_j^2<\sigma\Big) =: B_1+B_2,$$

where $B_1\leqslant \|f\|_\gamma^2\omega_{k_\varepsilon^*}\gamma_{k_\varepsilon^*}^{-1}\leqslant r\chi_\varepsilon$ because $f\in\mathcal{F}_\gamma^r$. Moreover, $B_2\leqslant 4dr\kappa_\sigma$ using Lemma A.1 (ii). The result of the theorem follows now by combination of the decomposition (A.5) and the estimates of A_1, A_2, B_1 and B_2. □

A.2. Adaptive estimation (Section 3)

The proofs in this section use the Lemmas A.3– A.6 from the auxiliary results section A.3 below.

Proof of Proposition 3.3. Using the model equation $Y_j=[g]_j+\sqrt{\varepsilon}\,\xi_j$, we have for all $t\in\mathcal{S}_k$ that

$$[\widehat{f}_k-f_k]_j = \frac{\sqrt{\varepsilon}\,\xi_j}{a_j} + \left(\frac{1}{X_j}\mathbf{1}_{[X_j^2\geqslant\sigma]}-\frac{1}{a_j}\right)\sqrt{\varepsilon}\,\xi_j + \left(\frac{1}{X_j}\mathbf{1}_{[X_j^2\geqslant\sigma]}-\frac{1}{a_j}\right)[g]_j.$$

Thus, we may decompose the norm $\|\widehat{f}_k-f_k\|_\omega^2$ in three terms according to

$$\begin{aligned}\|\widehat{f}_k-f_k\|_\omega^2 &\leqslant 3\sum_{j=1}^k\frac{\omega_j}{a_j}\varepsilon\,\xi_j^2 + 3\sum_{j=1}^k\omega_j\left(\frac{1}{X_j}\mathbf{1}_{[X_j^2\geqslant\sigma]}-\frac{1}{a_j}\right)^2\varepsilon\,\xi_j^2 + 3\sum_{j=1}^k\omega_j\left(\frac{1}{X_j}\mathbf{1}_{[X_j^2\geqslant\sigma]}-\frac{1}{a_j}\right)^2[g]_j^2\\ &=: 3\left\{T_k^{(1)}+T_k^{(2)}+T_k^{(3)}\right\}.\end{aligned}$$

Define the event

$$\Omega_\sigma := \left\{\forall\, 0<j\leqslant M_\sigma^+ \;\middle|\; \left|\frac{1}{X_j}-\frac{1}{a_j}\right|\leqslant\frac{1}{2\,a_j}\quad\text{and}\quad X_j^2\geqslant\sigma\right\}.$$

Since $\mathbf{1}\{X_j^2\geqslant\sigma\}\mathbf{1}\{\Omega_\sigma\}=\mathbf{1}\{\Omega_\sigma\}$, it follows that for all $1\leqslant j\leqslant K_{\varepsilon,\sigma}^+$ we have

$$\left(\frac{a_j}{X_j}\mathbf{1}\{X_j^2\geqslant\sigma\}-1\right)^2\mathbf{1}\{\Omega_\sigma\} = a_j^2\,\mathbf{1}\{\Omega_\sigma\}\left|\frac{1}{X_j}-\frac{1}{a_j}\right|^2\leqslant\frac{1}{4}.$$

Hence, $T_k^{(2)}\mathbf{1}_{\Omega_\sigma}\leqslant\frac{1}{4}T_k^{(1)}$ for all $1\leqslant k\leqslant K_{\varepsilon,\sigma}^+$, and thus

$$\begin{aligned}\max_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^+}\left(\|\widehat{f}_k-f_k\|_\omega^2-\frac{1}{6}\,\mathrm{pen}_k^a\right)_+ &\leqslant 4\sum_{k=1}^{K_{\varepsilon,\sigma}^+}\left(\sum_{j=1}^k\frac{\omega_j}{a_j}\varepsilon\,\xi_j^2-2\delta_k\varepsilon\right)_+\\ &\quad + 3\max_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^+}T_k^{(2)}\mathbf{1}_{\Omega_\sigma^c}+3\max_{1\leqslant k\leqslant K_{\varepsilon,\sigma}^+}T_k^{(3)}.\end{aligned}$$

Keeping in mind that $\mathbf{P}[\Omega_\sigma^c]\leqslant C(d)\sigma^2$ by virtue of Lemma A.6, the result follows immediately using Lemmas A.3, A.4, and A.5 below. □

Proof of Proposition 3.5. Let $\breve{f}_k := \sum_{1\leqslant j\leqslant k}[f]_j\mathbf{1}\{X_j^2\geqslant\sigma\}e_j$. It is easy to see that $\|\widehat{f}_k-\breve{f}_k\|^2\leqslant\|\widehat{f}_{k'}-\breve{f}_{k'}\|^2$ for all $k'\leqslant k$ and $\|\breve{f}_k-f\|^2\leqslant\|f\|^2$ for all $k\geqslant 1$. Thus, using that $1\leqslant\widehat{k}\leqslant(N_\varepsilon^\circ\wedge\sigma^{-1})$,

we can write

$$
\begin{aligned}
\mathbb{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\} &\leqslant 2\{\mathbb{E}\|\widehat{f}_{\widehat{k}} - \breve{f}_{\widehat{k}}\|_\omega^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\} + \mathbb{E}\|\breve{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\}\} \\
&\leqslant 2\Big\{\mathbb{E}\|\widehat{f}_{(N_\varepsilon^\circ \wedge \lfloor \sigma^{-1}\rfloor)} - \breve{f}_{(N_\varepsilon^\circ \wedge \lfloor \sigma^{-1}\rfloor)}\|_\omega^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\} + \|f\|_\omega^2\, \mathbf{P}[\mho_{\varepsilon,\sigma}^c]\Big\}.
\end{aligned}
$$

Moreover, using the Cauchy-Schwarz inequality, we conclude

$$
\begin{aligned}
&\mathbb{E}\|\widehat{f}_{(N_\varepsilon^\circ \wedge \lfloor \sigma^{-1}\rfloor)} - \breve{f}_{(N_\varepsilon^\circ \wedge \lfloor \sigma^{-1}\rfloor)}\|_\omega^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\} \\
&\qquad \leqslant 2\sigma^{-1} \sum_{1 \leqslant j \leqslant (N_\varepsilon^\circ \wedge \lfloor \sigma^{-1}\rfloor)} \omega_j \Big\{\mathbb{E}(Y_j - a_j[f]_j)^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\} + \mathbb{E}(a_j[f]_j - X_j[f]_j)^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\}\Big\} \\
&\qquad \leqslant 2\sigma^{-1}\Big\{ \sum_{1 \leqslant j \leqslant (N_\varepsilon^\circ \wedge \lfloor \sigma^{-1}\rfloor)} \omega_j \Big[\mathbb{E}\,(Y_j - [g]_j)^4\Big]^{1/2} \mathbf{P}[\mho_{\varepsilon,\sigma}^c]^{1/2} \\
&\qquad\qquad + \sum_{1 \leqslant < j \leqslant (N_\varepsilon^\circ \wedge \lfloor \sigma^{-1}\rfloor)} \omega_j [f_j]^2 [\mathbb{E}(X_j - a_j)^4]^{1/2} \mathbf{P}[\mho_{\varepsilon,\sigma}^c]^{1/2} \Big\} \\
&\qquad \leqslant 2\sqrt{3}\sigma^{-1}\Big\{(\sigma^{-1} \max_{1 \leqslant j \leqslant N_\varepsilon^\circ} \omega_j)\varepsilon + \sigma\|f\|_\omega^2\Big\}\, \mathbf{P}[\mho_{\varepsilon,\sigma}^c]^{1/2},
\end{aligned}
$$

which implies

$$
\mathbb{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\} \leqslant C\Big\{\Big(\sigma^{-2} + \|f\|_\omega^2\Big)\, \mathbf{P}[\mho_{\varepsilon,\sigma}^c]^{1/2} + \|f\|_\omega^2\, \mathbf{P}[\mho_{\varepsilon,\sigma}^c]\Big\}.
$$

Lemma A.6 below yields, for some $C > 0$ depending only on the class $\mathcal{A}_\lambda^d$,

$$
\mathbb{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}\{\mho_{\varepsilon,\sigma}^c\} \leqslant C\left\{\sigma + \|f\|_\omega^2 \sigma^6 + \|f\|_\omega^2 \sigma^{12}\right\}
$$

which completes the proof due to $f \in \mathcal{F}_\gamma^r$. $\square$

A.3. Auxiliary results

Lemma A.1 *For every $j \in \mathbb{N}$,*

(i) $R_j^I := \mathbb{E}\Bigg[\left(\frac{a_j}{X_j} - 1\right)^2 \mathbf{1}\{X_j^2 \geqslant \sigma\}\Bigg] \leqslant \min\Big\{1, \frac{8\sigma}{a_j^2}\Big\}$

(ii) $R_j^{II} := \mathbf{P}[X_j^2 < \sigma] \leqslant \min\Big\{1, \frac{4\sigma}{a_j^2}\Big\}$

(iii) $\mathbb{E}\left[\left(\frac{\mathbb{E}[X_j]}{X_j}\right)^2 \mathbf{1}\{X_j^2 \geqslant \sigma\}\right] \leqslant 4$

Proof. (i) It is easy to see that

$$
R_j^I = \mathbb{E}\left[\frac{|X_j - a_j|^2}{X_j^2}\ \mathbf{1}\{X_j^2 \geqslant \sigma\}\right] \leqslant \sigma^{-1}\, \mathbb{V}\mathrm{ar}(X_j) = 1. \tag{A.6}
$$

On the other hand, using that $\mathbb{E}[(X_j - a_j)^4] = 3\sigma^2$, we obtain

$$
\begin{aligned}
R_j^I &\leqslant \mathbb{E}\left[\frac{(X_j - a_j)^2}{X_j^2}\ \mathbf{1}\{X_j^2 \geqslant \sigma\}\ 2\Big\{\frac{(X_j - a_j)^2}{a_j^2} + \frac{X_j^2}{a_j^2}\Big\}\right] \\
&\leqslant \frac{2\,\mathbb{E}[(X_j - a_j)^4]}{\sigma a_j^2} + \frac{2\ \mathbb{V}\mathrm{ar}(X_j)}{a_j^2} = \frac{8\sigma}{a^2}.
\end{aligned}
$$

Combining with (A.6) gives $R_j^I \leqslant \min\left\{1, \frac{8\sigma}{a_j^2}\right\}$, which completes the proof of (i).
(ii) Trivially, $R_j^{II} \leqslant 1$. If $1 \leqslant 4\sigma/a_j^2$, then obviously $R_j^{II} \leqslant \min\left\{1, \frac{4\sigma}{a_j^2}\right\}$. Otherwise, we have $\sigma < a_j^2/4$ and hence, using Tchebychev's inequality,

$$R_j^{II} \leqslant \mathbf{P}[|X_j - a_j| > |a_j|/2] \leqslant \frac{4\ \mathbb{V}\mathrm{ar}(X_j)}{a_j^2} \leqslant \min\left\{1, \frac{4\sigma}{a_j^2}\right\},$$

where we have used that $\mathbb{V}\mathrm{ar}(X_j) = \sigma$ for all j.
(iii) $\mathbb{E}\left[\left(\frac{\mathbb{E}[X_j]}{X_j}\right)^2 \mathbf{1}\{X_j^2 \geqslant \sigma\}\right] \leqslant 2\mathbb{E}\left[\left(\frac{X_j - \mathbb{E}[X_j]}{X_j}\right)^2 \mathbf{1}\{X_j^2 \geqslant \sigma\} + \mathbf{1}\{X_j^2 \geqslant \sigma\}\right] \leqslant 4.$ □

Lemma A.2 *Under Assumption 2.1, we have that*

(i) $\varepsilon\delta_{N_\varepsilon^+} \leqslant 32\, d^2$ *for all* $\varepsilon \in (0,1)$,

and there is a $\sigma_0 \in (0,1)$ *such that for all* $\sigma < \sigma_0$, *we have*

(ii) $\min_{1\leqslant j\leqslant M_\sigma^+} a_j^2 \geqslant 3\sigma$.

Proof. (i) For $N_\varepsilon^+ = 0$, we have $\delta_{N_\varepsilon^+} = 0$ and there is nothing to show. If $0 < N_\varepsilon^+ \leqslant n$, one can show that $\omega^+_{N_\varepsilon^+}/\lambda_{N_\varepsilon^+} \leqslant 4d/(\varepsilon N_\varepsilon^+|\log\varepsilon|)$, which we use in the following computation:

$$\begin{aligned}
\delta_{N_\varepsilon^+} &= N_\varepsilon^+ \frac{\omega^+_{N_\varepsilon^+}}{\lambda_{N_\varepsilon^+}} \frac{\log((\omega^+_{N_\varepsilon^+}/\lambda_{N_\varepsilon^+}) \vee (N_\varepsilon^+ + 2))}{\log(N_\varepsilon^+ + 2)} \leqslant \frac{4d}{\varepsilon|\log\varepsilon|} \frac{\log\left(\frac{4d}{N_\varepsilon^+\varepsilon|\log\varepsilon|} \vee (N_\varepsilon^+ + 2)\right)}{\log(N_\varepsilon^+ + 2)} \\
&\leqslant \varepsilon^{-1} \begin{cases} 4d & (\log(\varepsilon^{-1} + 2) \geqslant 4d) \\ 4d(4d + \log(4d))/(\log(\varepsilon^{-1} + 2)) & (\text{otherwise}), \end{cases}
\end{aligned}$$

which implies $\varepsilon\delta_{N_\varepsilon^+} \leqslant 4d(4d + \log(4d)) \leqslant 32d^2$ for all $\varepsilon \in (0,1)$.
(ii) We have that

$$\min_{1\leqslant j\leqslant M_\sigma^+} a_j^2 \geqslant \min_{1\leqslant j\leqslant M_\sigma^+} \frac{\lambda_j}{d} \geqslant \frac{\sigma^{1-v_\sigma}}{4d^2} \geqslant 3\sigma,$$

where the last step holds for sufficiently small σ as some algebra shows. □

Lemma A.3 *We have that*

$$\sum_{k=1}^{K_{\varepsilon,\sigma}^+} \mathbb{E}\left(\sum_{j=1}^k \frac{\omega_j}{a_j}\varepsilon\xi_j^2 - 2\,\delta_k^a\varepsilon\right)_+ \leqslant 6720\,\varepsilon.$$

Proof. Representing the expectation of the positive random variable by the integral over its tail probabilities and using $\delta_k^a \geqslant \sum_{j=1}^k(\omega_j/a_j^2)$, we may write

$$\begin{aligned}
\sum_{k=1}^{K_{\varepsilon,\sigma}^+} \mathbb{E}\left(\sum_{j=1}^k \frac{\omega_j}{a_j}\varepsilon\xi_j^2 - 2\,\delta_k^a\varepsilon\right)_+ &\leqslant \sum_{k=1}^{K_{\varepsilon,\sigma}^+} \int_0^\infty \mathbf{P}\left[\sum_{j=1}^k \frac{\varepsilon\omega_j}{a_j^2}(\xi_j^2 - 1) \geqslant x + 2\varepsilon\delta_k^a - \varepsilon\sum_{j=1}^k \frac{\omega_j}{a_j^2}\right] dx \\
&\leqslant \sum_{k=1}^{K_{\varepsilon,\sigma}^+} \int_0^\infty \mathbf{P}\left[\sum_{j=1}^k \frac{\varepsilon\omega_j}{a_j^2}(\xi_j^2 - 1) \geqslant x + \varepsilon\delta_k^a\right] dx
\end{aligned}$$

Define $\rho_k := (\varepsilon\omega_k)/a_k^2$, $H_k := 4\varepsilon\Delta_k^a$, and $B_k := 2\varepsilon^2 \sum_{j=1}^k \omega_j^2/a_j^4$. It can be shown (see proof of Proposition A.1 in Dahlhaus and Polonik (2006)) that for all $1 \leqslant k' \leqslant k$ and $m \geqslant 2$, we have

$$\Big|\mathbb{E}\Big[\Big(\frac{\varepsilon\omega_{k'}}{a_{k'}^2}(\xi_{k'}^2 - 1)\Big)^m\Big]\Big| \leqslant m!\, \rho_{k'}^2\, H_k^{m-2}.$$

Hence, the assumption of Theorem 2.8 from Petrov (1995) is satisfied and splitting up the integral, we get the following bound:

$$\begin{aligned}
&\sum_{k=1}^{K_{\varepsilon,\sigma}^+} \mathbb{E}\Big(\sum_{j=1}^k \frac{\omega_j}{a_j}\varepsilon\xi_j^2 - 2\,\delta_k^a\varepsilon\Big)_+ \\
&\qquad \leqslant \sum_{k=1}^{K_{\varepsilon,\sigma}^+} \int_0^{B_k/H_k - \varepsilon\delta_k^a} \exp\Big(-\frac{(x+\varepsilon\delta_k^a)^2}{4B_k}\Big)dx + \int_{B_k/H_k-\varepsilon\delta_k^a}^{\infty} \exp\Big(-\frac{x+\varepsilon\delta_k^a}{4H_k}\Big)dx
\end{aligned}$$

The second integral is equal to $4H_k\exp(-B_k/(4H_k^2))$. Some computation shows that the first one is bounded from above by $4H_k\big[\exp\big(-\varepsilon^2(\delta_k^a)^2/(4B_k)\big) - \exp\big(-B_k/(4H_k^2)\big)\big]$. Thus, the two identical terms cancel, and we get

$$\sum_{k=1}^{K_{\varepsilon,\sigma}^+} \mathbb{E}\Big(\sum_{j=1}^k \frac{\omega_j}{a_j}\varepsilon\xi_j^2 - 2\,\delta_k^a\varepsilon\Big)_+ \leqslant 16\,\varepsilon \sum_{k=1}^{K_{\varepsilon,\sigma}^+} \Delta_k^a \exp\Big(-\frac{(\delta_k^a)^2}{8k(\Delta_k^a)^2}\Big).$$

To complete the proof, we bound the sum on the right hand side as follows,

$$\begin{aligned}
\sum_{k=1}^{K_{\varepsilon,\sigma}^+} \Delta_k^a \exp\Big(-\frac{(\delta_k^a)^2}{8k(\Delta_k^a)^2}\Big) &\leqslant \sum_{k=1}^{\infty} \exp\Big(-\log(\Delta_k^a \vee (k+2))\Big[\frac{k}{8\log(k+2)} - 1\Big]\Big) \\
&\leqslant e\sum_{k=1}^{\infty} \exp\Big(-\frac{k}{8\log(k+2)}\Big) \leqslant e\sum_{k=1}^{\infty}\exp\Big(-\frac{\sqrt{k}}{8\log(3)}\Big) \\
&\leqslant e\int_0^\infty \exp\Big(-\frac{\sqrt{x}}{8\log(3)}\Big)dx = 128\log^2(3)\,e,
\end{aligned}$$

where we have used $\log(k+2) \leqslant \log(3)\sqrt{k}$ for all $k \geqslant 1$. ☐

Lemma A.4 *For every $k \in \mathbb{N}$ and $\sigma \in (0,1)$,*

$$\mathbb{E}\Big[\sum_{j=1}^k \omega_j [g]_j^2 \Big(\frac{1}{X_j}\mathbf{1}_{[X_j \geqslant \sigma]} - \frac{1}{a_j}\Big)^2\Big] \leqslant 8\; d\; r\; \kappa_\sigma(\gamma, \lambda, \omega).$$

Proof. Firstly, as $f \in \mathcal{F}_\gamma^r$, it is easily seen that

$$\mathbb{E}\Big[\sum_{j=1}^k \omega_j [g]_j^2 \Big(\frac{1}{X_j}\mathbf{1}_{[X_j \geqslant \sigma]} - \frac{1}{a_j}\Big)^2\Big] \leqslant r \max_{1\leqslant j\leqslant k} \frac{\omega_j}{\gamma_j}\,\mathbb{E}[|R_j|^2],$$

where R_j is defined as

$$R_j := \Big(\frac{a_j}{X_j}\mathbf{1}\{X_j^2 \geqslant \sigma^2\} - 1\Big). \tag{A.7}$$

In view of the definition of κ_σ in Theorem 2.3, the result follows from $\mathbb{E}[|R_j|^2] \leqslant d \min\left\{1, \frac{8\sigma}{\lambda_j}\right\}$, which is a consequence of the decomposition

$$\mathbb{E}|R_j|^2 = \mathbb{E}\left[\left(\frac{a_j}{X_j} - 1\right)^2 \mathbf{1}\{X_j^2 \geqslant \sigma\}\right] + \mathbf{P}[X_j^2 < \sigma] \tag{A.8}$$

and Lemma A.1. □

Lemma A.5 *We have that*

$$\mathbb{E}\left[\sum_{j=1}^{K^+_{\varepsilon,\sigma}} \omega_j \left(\frac{1}{X_j}\mathbf{1}_{[X_j \geqslant \sigma]} - \frac{1}{a_j}\right)^2 \varepsilon \xi_j^2 \mathbf{1}_{\Omega_\sigma^c}\right] \leqslant 64\, d^3 (\mathbf{P}[\Omega_\sigma^c])^{1/2}.$$

Proof. Given R_j from (A.7), we begin our proof observing that

$$\mathbb{E}\left[\sum_{j=1}^{K^+_{\varepsilon,\sigma}} \omega_j \left(\frac{1}{X_j}\mathbf{1}_{[X_j \geqslant \sigma]} - \frac{1}{a_j}\right)^2 \sqrt{\varepsilon} \xi_j^2 \mathbf{1}_{\Omega_\sigma^c}\right] \leqslant \varepsilon \sum_{j=1}^{K^+_{\varepsilon,\sigma}} \frac{\omega_j}{a_j^2}\, \mathbb{E}[|R_j|^2 \mathbf{1}_{\Omega_\sigma^c}],$$

where we have used the independence of X and Y and $\mathbb{V}\mathrm{ar}(Y_j) = \varepsilon$. Since $d\delta_k^\lambda \geqslant \sum_{j=1}^k \frac{\omega_j}{a_j^2}$ for all $A \in \mathcal{A}_\lambda^d$, the Cauchy-Schwarz inequality yields

$$\mathbb{E}\left[\sum_{j=1}^{K^+_{\varepsilon,\sigma}} \omega_j \left(\frac{1}{X_j}\mathbf{1}_{[X_j \geqslant \sigma]} - \frac{1}{a_j}\right)^2 \varepsilon \xi_j^2 \mathbf{1}_{\Omega_\sigma^c}\right] \leqslant\ d\, (\mathbf{P}[\Omega_\sigma^c])^{1/2}\, \varepsilon \delta^\lambda_{N_\varepsilon^+} \max_{0<j\leqslant N_\varepsilon^+} (\mathbb{E}[|R_j|^4])^{1/2}.$$

Proceeding analogously to (A.6) and (A.8), one can show that $\mathbb{E}[|R_j|^4] \leqslant 4$. The result follows then using the definition of N_ε^+. □

Lemma A.6 *For $k \in \mathbb{N}$, define the events*

$$\widetilde{\Omega}_k := \left\{\left|\frac{X_j}{a_j} - 1\right| \leqslant \frac{1}{3} \quad \forall\, 1 \leqslant j \leqslant k\right\}$$

and suppose that Assumption 2.1 holds. For all $\varepsilon, \sigma \in (0,1)$, we have

(i) $\Omega_\sigma \subseteq \{\mathrm{pen}_k^+ \leqslant \widehat{\mathrm{pen}}_k \leqslant 30\, \mathrm{pen}_k^+ \quad \forall\, 1 \leqslant k \leqslant K^+_{\varepsilon,\sigma}\}$,

(ii) $\widetilde{\Omega}_{M_\sigma^+ + 1} \subseteq \{K^-_{\varepsilon,\sigma} \leqslant \widehat{K}_{\varepsilon,\sigma} \leqslant K^+_{\varepsilon,\sigma}\}$,

(iii) $\mathbf{P}[\widetilde{\Omega}^c_{M_\sigma^+}] \leqslant C(d)\, \sigma^2$ *and* $\mathbf{P}[\Omega_\sigma^c] \leqslant C(d)\, \sigma^2$.

If additionally condition (3.6) holds, then

(iv) $\mathbf{P}[\mho^c_{\varepsilon,\sigma}] \leqslant C(\lambda, d)\sigma^6$.

Proof. Consider (i). Notice first that $\delta_k^a \leqslant \delta_k^\lambda\, d\, \zeta_d$ for all $k \geqslant 1$ with $\zeta_d := (\log(3d))/(\log 3)$. Observe that on Ω_σ we have $(1/2)\Delta_k^a \leqslant \Delta_k^X \leqslant (3/2)\Delta_k^a$ for all $1 \leqslant k \leqslant \widetilde{M}_\sigma$ and hence $(1/2)[\Delta_k^a \vee (k+2)] \leqslant [\Delta_k^X \vee (k+2)] \leqslant (3/2)[\Delta_k^a \vee (k+2)]$, which implies

$$\begin{aligned}(1/2)k\Delta_k^a\Big(\frac{\log[\Delta_k^a \vee (k+2)]}{\log(k+2)}\Big)\Big(1 - \frac{\log 2}{\log(k+2)}\frac{\log(k+2)}{\log(\Delta_k^a \vee [k+2])}\Big) \\ \leqslant \delta_k^X \leqslant (3/2)k\Delta_k^a\Big(\frac{\log(\Delta_k^a \vee [k+2])}{\log(k+2)}\Big)\Big(1 + \frac{\log 3/2}{\log(k+2)}\frac{\log(k+2)}{\log(\Delta_k^a \vee [k+2])}\Big).\end{aligned}$$

Using $\log(\Delta_k^a \vee (k+2))/\log(k+2) \geqslant 1$, we conclude from the last estimate that

$$\begin{aligned}\delta_k^a/10 \leqslant &(\log 3/2)/(2\log 3)\delta_k^a \leqslant (1/2)\delta_k^a[1-(\log 2)/\log(k+2)] \leqslant \delta_k^X \\ &\leqslant (3/2)\delta_k^a[1+(\log 3/2)/\log(k+2)] \leqslant 3\delta_k^a.\end{aligned}$$

It follows that on Ω_σ we have $\mathrm{pen}_k^+ \leqslant \widehat{\mathrm{pen}}_k \leqslant 30\,\mathrm{pen}_k^+$ for all $1 \leqslant k \leqslant M_\sigma^+$ as desired.

Proof of (ii). Denoting by X the random sequence $(X_j)_{j\geqslant 1}$, define sequences $N_\varepsilon^- := N_\varepsilon^{\sqrt{\lambda/(4d)}}$, $M_\sigma^- := M_\sigma^{\sqrt{\lambda/(4d)}}$ and $\widehat{N}_\varepsilon := N_\varepsilon^X$, $\widehat{M}_\sigma := M_\sigma^X$. Note that by definition, $K_{\varepsilon,\sigma}^- = N_\varepsilon^- \wedge M_\sigma^-$ and $\widehat{K}_{\varepsilon,\sigma} = \widehat{N}_\varepsilon \wedge \widehat{M}_\sigma$. Define further the events $\Omega_I := \{K_{\varepsilon,\sigma}^- > \widehat{K}_{\varepsilon,\sigma}\}$ and $\Omega_{II} := \{\widehat{K}_{\varepsilon,\sigma} > K_{\varepsilon,\sigma}^+\}$. Then we have $\{K_{\varepsilon,\sigma}^- \leqslant \widehat{K}_{\varepsilon,\sigma} \leqslant K_{\varepsilon,\sigma}^+\}^c = \Omega_I \cup \Omega_{II}$. Consider $\Omega_I = \{\widehat{N}_\varepsilon < K_{\varepsilon,\sigma}^-\} \cup \{\widehat{M}_\sigma < K_{\varepsilon,\sigma}^-\}$ first. By definition of N_ε^-, we have that $\min_{1\leqslant j\leqslant N_\varepsilon^-} \frac{a_j^2}{j\,\omega_j^+} \geqslant 4\,\varepsilon|\log\varepsilon|$, which implies, keeping in mind that $K_{\varepsilon,\sigma}^- \leqslant N_{\varepsilon,\sigma}^-$,

$$\begin{aligned}\{\widehat{N}_\varepsilon < K_{\varepsilon,\sigma}^-\} &\subset \left\{\exists\, 1 \leqslant j \leqslant K_{\varepsilon,\sigma}^- \;\middle|\; \frac{X_j^2}{j\,\omega_j^+} < \varepsilon|\log\varepsilon|\right\} \\ &\subset \bigcup_{1\leqslant j\leqslant K_{\varepsilon,\sigma}^-}\left\{\frac{X_j}{a_j} \leqslant \frac{1}{2}\right\} \subset \bigcup_{1\leqslant j\leqslant K_{\varepsilon,\sigma}^-}\left\{\left|\frac{X_j}{a_j}-1\right| \geqslant \frac{1}{2}\right\}.\end{aligned}$$

One can see that from $\min_{1\leqslant j\leqslant M_\sigma^-} a_j^2 \geqslant 4\sigma^{1-v_\sigma}$ it follows in the same way that

$$\left\{\widehat{M}_\sigma < K_{\varepsilon,\sigma}^-\right\} \subset \bigcup_{1\leqslant j\leqslant K_{\varepsilon,\sigma}^-}\left\{\left|\frac{X_j}{a_j}-1\right| \geqslant \frac{1}{2}\right\}.$$

Therefore, $\Omega_I \subseteq \bigcup_{1\leqslant j\leqslant M_\sigma^+}\left\{|X_j/a_j - 1| \geqslant 1/2\right\} \subseteq \widetilde{\Omega}_{M_\sigma^+ + 1}^c$, since $M_\sigma^- \leqslant M_\sigma^+$.
Consider $\Omega_{II} = \{\widehat{N}_\varepsilon > K_{\varepsilon,\sigma}^+\} \cap \{\widehat{M}_\sigma > K_{\varepsilon,\sigma}^+\}$. In case $K_{\varepsilon,\sigma}^+ = N_\varepsilon^+$, note that by definition of N_ε^+, we have $\varepsilon|\log\varepsilon|/4 \geqslant \frac{a_{N_\varepsilon^+ +1}^2}{(N_\varepsilon^+ +1)\,\omega_{N_\varepsilon^+ +1}^+}$, such that

$$\begin{aligned}\Omega_{II} \subseteq \{\widehat{N}_\varepsilon > N_\varepsilon^+\} &\subset \left\{\forall 1 \leqslant j \leqslant N_\varepsilon^+ + 1 \;\middle|\; \frac{X_j^2}{j\,\omega_j^+} \geqslant \varepsilon|\log\varepsilon|\right\} \\ &\subset \left\{\frac{X_{N_\varepsilon^+ +1}}{a_{N_\varepsilon^+ +1}} \geqslant 2\right\} \subset \left\{\left|\frac{X_{N_\varepsilon^+ +1}}{a_{N_\varepsilon^+ +1}} - 1\right| \geqslant 1\right\}.\end{aligned}$$

In case $K_{\varepsilon,\sigma}^+ = M_\sigma^+$, it follows analogously from $\sigma^{1-v_\sigma} \geqslant 4\max_{j\geqslant M_\sigma^+ +1} a_j^2$ that

$$\Omega_{II} \subset \{\widehat{M}_\sigma > M_\sigma^+\} \subset \left\{|X_{M_\sigma^+ +1}/a_{M_\sigma^+ +1} - 1| \geqslant 1\right\}.$$

Therefore, we have $\Omega_{II} \subseteq \left\{|X_{K_{\varepsilon,\sigma}^+ +1}/a_{K_{\varepsilon,\sigma}^+ +1} - 1| \geqslant 1\right\} \subseteq \widetilde{\Omega}_{M_\sigma^+ +1}^c$ and (ii) is shown.

Proof of (iii). For $Z \sim \mathcal{N}(0,1)$ and $z \geqslant 0$, one has $\mathbf{P}[Z > z] \leqslant (2\pi z^2)^{-1/2}\exp(-z^2/2)$. Hence, there is a constant $C(d)$ depending on d such that for every $1 \leqslant j \leqslant M_\sigma^+$,

$$\mathbf{P}[|X_j/a_j - 1| > 1/3] \leqslant C(d)\left(\frac{\sigma}{\lambda_{M_\sigma^+}}\right)^{1/2}\exp\left(-\frac{\lambda_{M_\sigma^+}}{18\sigma d}\right).$$

Consequently, as $M_\sigma^+ \leqslant \sigma^{-1}$ and $\lambda_{M_\sigma^+} > \sigma^{1-v_\sigma}/(4d)$, we have

$$\mathbf{P}[\widetilde{\Omega}^c_{M_\sigma^+}] \leqslant C(d)\sigma^{2-v_\sigma} \exp\left(- \frac{\sigma^{-v_\sigma}}{72d^2}\right)$$

which implies $\mathbf{P}[\widetilde{\Omega}^c_{M_\sigma^+}] \leqslant C(d)\, \sigma^2$ using that $\sigma^{v_\sigma}\, |\log \sigma| \to 0$ as $\sigma \to 0$. As for the second assertion in (iii), we distinguish the cases $\sigma \leqslant \sigma_0$ and $\sigma > \sigma_0$, where σ_0 is the constant from Lemma A.2 (ii) depending only on d. The assertion is trivial for $\sigma > \sigma_0$ (keeping in mind that $\mathbf{P}[\Omega_\sigma^c] \leqslant \sigma_0^{-2}\sigma^2$). Consider the case $\sigma \leqslant \sigma_0$, where $a_j^2 \geqslant 3\sigma$ for all $1 \leqslant j \leqslant M_\sigma^+$ due to Lemma A.2 (ii). This yields for the complement of Ω_σ

$$\Omega_\sigma^c = \left\{ \exists\, 1 \leqslant j \leqslant M_\sigma^+ \quad \middle| \quad \left| \frac{a_j}{X_j} - 1 \right| > \frac{1}{2} \quad \text{or} \quad X_j^2 < \sigma \right\} \subseteq \left\{ \exists\, 1 \leqslant j \leqslant M_\sigma^+ \;\middle|\; \left| \frac{X_j}{a_j} - 1 \right| > \frac{1}{3} \right\} = \widetilde{\Omega}^c_{M_\sigma^+}.$$

It follows with assertion (ii) that $\mho^c_{\varepsilon,\sigma} \subseteq \widetilde{\Omega}^c_{M_\sigma^+}$ for all $\sigma \leqslant \sigma_0$, implying the second assertion of (iii).

Proof of (iv). Following the proof of (iii) and using that $M_\sigma^+ + 1 \leqslant \sigma^{-1}$, we obtain

$$\mathbf{P}[\widetilde{\Omega}^c_{M_\sigma^+ + 1}] \leqslant C(d)(\sigma \lambda_{M_\sigma^+ + 1})^{-1/2} \exp\left(- \frac{\lambda_{M_\sigma^+ + 1}}{18\sigma d}\right). \tag{A.9}$$

Note that $\widetilde{\Omega}_{M_\sigma^+ + 1} \subseteq \Omega_\sigma$, since trivially $\widetilde{\Omega}_{M_\sigma^+ + 1} \subseteq \widetilde{\Omega}_{M_\sigma^+}$. Thus, (A.9) implies assertion (iv) by virtue of condition (3.6). $\square$

References

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413.

Cavalier, L., Golubev, G., Picard, D., and Tsybakov, A. (2002). Oracle inequalities for inverse problems. *Ann. Stat.*, 30:843–874.

Cavalier, L. and Hengartner, N. W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems*, 21:1345–1361.

Dahlhaus, R. and Polonik, W. (2006). Nonparametric quasi-maximum likelihood estimation for Gaussian locally stationary processes. *Ann. Stat.*, 34:2790–2824.

Efromovich, S. (1997). Density estimation for the case of supersmooth measurement error. *Journal of the American Statistical Association*, 92:526–535.

Ermakov, M. (1990). On optimal solutions of the deconvolution problem. *Inverse Probl.*, 6(5):863–872.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19:1257–1272.

Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Stat.*, 39(3):1608–1632.

Hoffmann, M. and Reiss, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, 36:310–336.

Johnstone, I. M. and Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Stat.*, 18(1):251–280.

Mair, B. A. and Ruymgaart, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM Journal on Applied Mathematics*, 56(5):1424–1444.

Mathé, P. and Pereverzev, S. V. (2001). Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. *SIAM J. Numer. Anal.*, 38(6):1999–2021.

Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7:307–330.

Petrov, V. V. (1995). *Limit theorems of probability theory. Sequences of independent random variables.* Oxford Studies in Probability. Clarendon Press., Oxford, 4. edition.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21:169–184.